6. 모델을 통한 생존, 사망 여부 예측

다음으로 4가지 모델을 활용하여 말의 사망과 생존을 예측해보고, 각 모델의 정분류율, 민감도와 특이도를 비교해 보도록 하겠다.

첨언) 잠재층 분석에서 살펴봤듯이 안락사와 생존 또는 사망 간에 특성은 다소 애매하다고 판단된다. 즉, 말이 사망할 것으로 예상하여 안락사를 시켰는지, 아니면 상품가치가 떨어져서 안락사를 시킨 것인지 판단 할 수 없었다. 따라서 문제를 좀더 명확히 하기 위해 앞으로의 데이터 분석에서 안락사한 경우의 데이터는 쓰지 않기로 했다. 따라서 관측 개체의 개수 255개만을 고려한다. (먼저 outcome이 안락사였던 경우를 데이터에서 제외하고 다시 이전과 같은 방법으로 imputation을 시행했다.)

6-1 Support Vector Machine (SVM)

대표적 지도학습머신인 서포트벡터머신은 classification에 유용하다. 따라서 outcome을 반응변수로, 나머지 변수를 설명변수로 처리하여 서포트벡터머신을 통해 어떤 말이 사망할지 또는 생존할지 분류를 하고자 한다. R의 "e1071" 패키지의 svm() 함수를 사용하여 분석을 진행하였다.

이때 전체 데이터의 70%는 훈련데이터로, 나머지 30%의 데이터는 모델의 성능을 확인하기 위한 목적으로 테스트 데이터로서 사용하였다. 먼저, 서포트벡터머신의 초모수인 gamma와 cost에 따라 모델의 성능의 차이가 많이 달라질 수 있기 때문에 다음과 같이 tune 함수를 사용하여 SVM의 최적 파라미터를 구하도록 한다. 이 때 자료의 불균형을 해소해 주고자 class.weights 옵션을 이용하여 클래스별 weight를 주었다. 그 결과는 다음과 같다.

```
tune.svm <- tune(svm, outcome ~., data=train, ranges = list(gamma=c(0.05,0.1,0.5,0.7), cost =
c(1,5,10,15,20,30,40,50,60,70,80), kernel= c("radial","linear","polynomial","sigmoid")),class.weights = c("1"=0.3, "2"=0.7))

summary(tune.svm)
```

```
Parameter tuning of 'svm':
- sampling method: 10-fold cross validation
- best parameters:
- best performance: 0.147619
- Detailed performance results:
```

gamma= 0.05, C = 1, kernel = "linear" 을 선택했을 때 가장 좋은 퍼포먼스를 보였다. 성능이 가장 좋았을 때 오분류율 14.76% 였으므로, 85.23%의 훈련데이터를 정확히 분류해냈다는 것을 알 수 있었다. 모델의 성능 을 검증하기 위해 최적의 파라미터로 훈련데이터를 통해 모델을 적합 시

키고 테스트 데이터를 분 류해 보도록 하겠다. 분류 결과는 다음과 같았다.

```
svm.model.1 <- svm(outcome ~., data = horse.1 , gamma = 0.05, cost = 1,kernel = "linear",class.weights = c("1"=0.3, "2"=0.7))

svm.predict.2 <- predict(svm.model.1, newdata = horse.2)

addmargins(table(test$outcome, svm.predict.2))
```

```
     1  2  Sum
1   28 10  38
2    4  9  13
Sum 32 19  51
```

제대로 분류된 값은 51개의 관측값 중 총 28 + 9 = 37였다. 따라서 37/51 = 72.55%를 제대로 분류해냈으며 오분류율 13.73% 였다. 민감도는 약 28/38 = 73.68% 였으며, 특이도는 약 9/13 = 69.23%였다.

6-2나무 알고리듬

마찬가지로 outcome을 반응변수로, 나머지 변수를 설명변수로 처리하여 나무 알고리듬을 통해 어떤 말이 사망할지 또는 생존할지 분류를 하고자 한다. 훈련데이터에 R의 "rpart" 패키지의 rpart() 함수를 사용하여 모델을 만들었고 테스트 데이터를 통해 모델을 평가했다. 자료에 불균형이 있었기에 훈련 데이터에서 outcome이 사망이었던 경우를 두배로 부풀렸다. 불순도 평가지수는 gini로 설정했다. 다음으로 최선의 모수를 찾기위해 다음과 같이 튜닝하였으며, 최선의 모수로 훈련 데이터를 학습시키고, 테스트 데이터를 예측한 결과는 다음과 같았다.

```
#자료의 불균형을 맞추기 위해 새롭게 data.balanced 데이터를 생성함
data.train.live = data.train[data.train$outcome==1,]
data.train.dead = data.train[data.train$outcome==2,]
data.balanced <- rbind(data.train.live, data.train.dead)
# 튜닝
horse.tree <- formula(outcome ~ .,data= data.balanced)
horse.tune = tune.rpart(horse.tree, data= data.balanced, maxdepth = c(3,4,5,6,7) , cp = c(0.01,0.02,0.03,0.04))
```

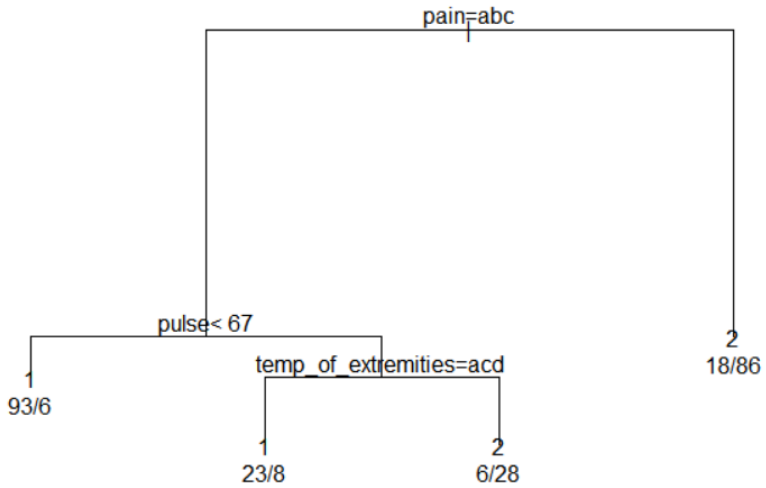
```
Parameter tuning of ‘rpart.wrapper’:
- sampling method: 10-fold cross validation
- best parameters:
- best performance: 0.1497151
- Detailed performance results:
```

가장 성능이 좋았을 때 오 분류율은 14.97% 였고 정 분류율은 85.03 이었음을 알 수 이었다. 최선의 파라미터는 cp =0.03, maxdepth=3 이었다. 다음 코드를 통해 훈련데이터를 적합시키고 테스트 데이터를 예측해 보았다.

```
data.tree.1 <- rpart(outcome ~ ., data=data.balanced, control = rpart.control(cp=0.03),
maxdepth=3)
addmargins(table(data.test$outcome,test))
```

```
     1  2 Sum
1   29  9  38
2    6  7  13
Sum 35 16  51
```

제대로 분류된 값은 51개의 관측값 중 총 29+7 = 36 이었다. 따라서 36/51 = 70.58%를 제대로 분류해냈으며 오분류율 29.41% 였다. 민감도는 약 29/38 = 76.31% 였으며 특이도는 약 7/13 = 53.85%였다. 추가적으로 분류에 사용된 알고리듬은 다음과 같이 플롯을 그릴 수 있었다.



<그림 1>

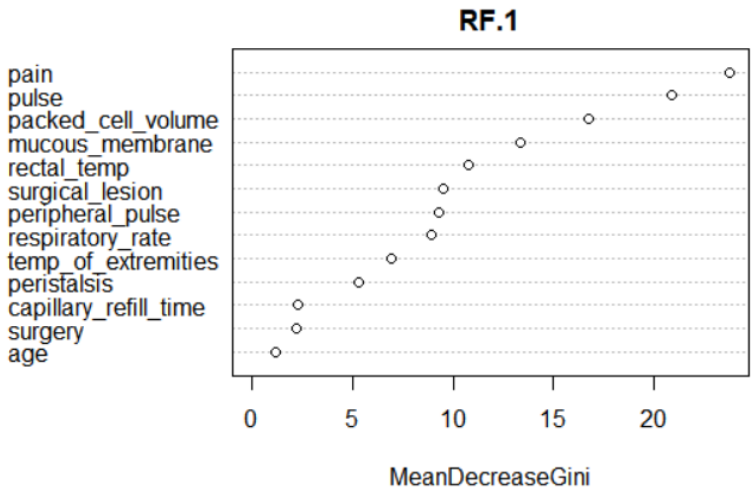
Pain이 1,2,3번째 범주가 아닐 경우(즉, 심한 고통인 경우) outcome이 2(사망)으로 할당되었으며, 1,2,3 번째 범주일 경우(즉 고통이 심하지 않은 경우)는 다음 질문 pulse 가 67 이하인가? 라는 질문을 받는다. 여기서 대답이 yes 일 시에는 outcome 이 1(생존) 으로 분류되며 no일시에는 temp_of_extremities가 1,3,4 번째 범주인가란 질문으로 넘어간다. 이 때 대답이 yes일시에는 outcome이 1(생존)으로 분류되며, no 일시에는 outcome이 2(사망)으로 분류된다는 것을 알 수 있었다.

6-3랜덤 포레스트

랜덤 포레스트는 배깅의 일종으로, 나무모형을 확장하여 부표본과 설명변수를 임의 추출하여 모델을 적합시킨다. CART에서 특정 변수에 대해 민감할 수 있는 문제를 해결하여 좀 더 로버스트하다고 할 수 있다. 훈련 데이터를 적합시킨 코드와 결과는 다음과 같다. 여기서도 마찬가지로 outcome 이 사망이었던 경우를 두배로 불려주어 자료 불균형을 해소해 주었다.

```
data.train.live = data.train[data.train$outcome==1,]

data.train.dead = data.train[data.train$outcome==2,]

data.balanced <- rbind(data.train.live, data.train.dead,data.train.dead)

library(randomForest)

RF.1 <- randomForest(outcome ~. , data=data.balanced, var.importance=T, ntree=10000)

RF.1

varImpPlot(RF.1)
```

```
Call:
 randomForest(formula = outcome ~ ., data = data.balanced, var.importance =
T,     ntree = 10000)
               Type of random forest: classification
                     Number of trees: 10000
No. of variables tried at each split: 3

       OOB estimate of  error rate: 8.21%
Confusion matrix:
    1   2 class.error
1 118  22   0.1571429
2   0 128   0.0000000
```



<그림 2>

각 마디에서는 3개의 변수를 임의 추출했으며, 10000 번 반복으로 얻은 통합모형의 오류율은 8.21%였다. 변수 중요도는 MeanDecreaseGini 기준에 따라 pain이 가장 컸으며, pulse, packed_cell_volume, mucous_membrane도 큰 편이었으며, age와 surgery, capillary_refill_time은 분류 하는데 있어 중요도가 떨어졌다. 다음으로 적합된 모델을 통해 테스트 데이터를 예측해 보았다.

```
test = predict(RF.1,newdata=data.test, type="class")

addmargins(table(data.test$outcome,test))
```

```
     1  2 Sum
1   27 11  38
2    4  9  13
Sum 31 20  51
```

제대로 분류된 값은 51개의 관측값 중 총 27+9 = 36 이었다. 따라서 36/51 = 70.58%를 제대로 분류해냈으며 오분류율 29.41% 였다. 민감도는 약 27/38 = 71.05% 였으며 특이도는 약 9/13 = 69.23%였다.

6-4로지스틱 회귀

마지막으로 일반화선형모형을 적합 시키도록 하겠다. 종속변수 outcome 이 이항형 반응 변수이므로 로지스틱 회귀에 적합 시킬 수 있을 것이다. 회귀모형에서는 변수의 개수가 너무 많으면 오히려 예측에 안 좋을 수 있으므로 stepwise 방법을 이용해 변수를 선택했다. (마찬가지로 자료 불균형을 해소한 데이터를 사용했다.)

```
data.balanced$outcome <- relevel(data.balanced$outcome, ref = "2")

horse.glm<-glm(outcome ~ . , family="binomial", data= data.balanced)

h<- step(horse.glm)

for_stepwise =h$formula

horse.glm_step<-glm(for_stepwise , family="binomial", data= data.balanced)

summary(horse.glm_step)
```

```
Call:
glm(formula = for_stepwise, family = "binomial", data = data.balanced)
```

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.86157  -0.42540   0.00006   0.26361   2.33250

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)        3.779892   1.940902   1.947 0.051476 .
surgery2           1.925582   0.781402   2.464 0.013729 *
pulse             -0.032855   0.009885  -3.324 0.000888 ***
respiratory_rate   0.026496   0.015760   1.681 0.092729 .
mucous_membrane2  -2.598659   0.810577  -3.206 0.001346 **
mucous_membrane3  -0.828590   0.767606  -1.079 0.280389
mucous_membrane4  -0.149638   0.875229  -0.171 0.864247
mucous_membrane5  -2.020788   1.052416  -1.920 0.054840 .
mucous_membrane6   0.453258   1.169343   0.388 0.698299
pain2              1.160055   1.031760   1.124 0.260866
pain3              2.232355   1.044941   2.136 0.032651 *
pain4             -2.834830   1.064977  -2.662 0.007771 **
pain5             -1.445054   0.966316  -1.495 0.134803
peristalsis2      16.302531 995.000408   0.016 0.986928
peristalsis3      -1.938731   1.003446  -1.932 0.053351 .
peristalsis4      -1.031905   1.013728  -1.018 0.308711
packed_cell_volume -0.045028   0.028089  -1.603 0.108924
surgical_lesion2   4.849474   1.026443   4.725 2.31e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 370.99  on 267  degrees of freedom
Residual deviance: 153.72  on 250  degrees of freedom
AIC: 189.72

Number of Fisher Scoring iterations: 16
```

변수를 선택한 결과 surgery, pulse, respiratory_rate, mucous_membrane, pain, aeristalsis, packed_cell_volume, surgical_lesion 변수가 선택되었다. 여기서의 성공확률은 outcome =1 (생존)인 경우이다. 유의한 p-value를 기록했던 변수에 대해서만 해석을 해보자면 다음과 같다(유의수준 5%이하). 먼저 수술을 받은 경우, 받지 않은 경우에 비해 생존할 확률이 높아졌다.(포함된 변수에 따라 회귀계수가 달라질 수 있어 이전 분석과는 해석에 차이가 난 것으로 예상한다. 정확한 관계는 후에 4장에서 밝히도록 하겠다.) Pulse가 높을수록 생존확률이 낮아졌다. 점막이 두번째 범주(bright pink)인 경우 첫번째 범주(normal pink)인 경우에 비해 생존할 확률이 낮아졌다. (나머지 범주들일 경우에도 normal pink 인 경우보다 생존 확률이 낮아졌다. 그러나 유의하지는 않았다.) Pain이 3범주(intermittent mild pain)인 경우 1범주(no pain)인 경우보다 생존확률이 높아졌지만 pain이 4범주(intermittent severe pain) 인 경우에는 생존확률이 낮아졌다. 또한 외과적 병변이 아닐 경우 생존확률이 높아졌다.

```
glm_pred_step=predict(horse.glm_step, data.test, type = "response")
prediction_step=ifelse(glm_pred_step>= 0.5, 1, 2)
addmargins(table(data.test$outcome, prediction_step))
```

```
     1  2  Sum
1   29  9   38
2    7  6   13
Sum 36 15  51
```

제대로 분류된 값은 51개의 관측 값 중 총 29+6 = 35 이었다. 따라서 35/51 = 69.23%를 제대로 분류해냈으며 오분류율 30.77% 였다. 민감도는 약 29/38 = 76.32% 였으며 특이도는 약 6/13 =46.15% 였다.

6-5 모델 비교

<표 3 : 모델에 따른 정분류율, 민감도, 특이도>

	정분류율(%)	민감도(%)	특이도(%)
SVM	72.55	73.68	69.23
나무 알고리듬	70.58	76.31	53.85
랜덤 포레스트	70.58	71.05	69.23
로지스틱회귀	69.23	76.32	46.15

4개의 알고리듬 모두 예측력은 차이가 많이 나지 않았다. 4개의 알고리듬 중에서 SVM의 예측력이 가장 뛰어났다. SVM이 변수가 많은 상황에서도 잘 작동하기 때문인 것으로 보인다. 하지만 SVM은 변수에 대한 영향력을 해석 하는데는 한계가 있기 때문에 변수에 대한 해석을 하고자 한다면 나무 알고리듬이나 랜덤 포레스트, 로지스틱 회귀를 이용해야 한다.

나무 알고리듬과 랜덤포레스트 모두에서 예측에 중요했던 변수는 pain, pulse 였다. 추가적으로 나무 알고리듬에서는 temp_of_extremities 변수도 중요했으며, 랜덤 포레스트에서는 packed_cell_volume, mucous_membrane 변수가 중요했다.

한편 민감도는 나무 알고리듬과 로지스틱회귀가 좋았으며, 특이도는 SVM과 랜덤포레스트가 높았다. 따라서 어떤 상황을 중시하느냐에 따라 알고리듬을 다르게 선택해야 할것이다. 예를들어 어떤 말이 사망할 것을 예측하고 안락사 하고자 하는 상황에서는 특이도가 높은 모델을 선택해야한다.

하지만 전반적으로 모델의 예측력이 좋은 편은 아니었으며 과적합의 문제가 나타났다.

7. 다양한 통계적 방법을 활용한 분석

Horse colic dataset을 활용하여 다양한 측면에서 통계적 분석을 수행할 수 있었다. 다음으로 수행한 결과는 모두 SAS base를 활용하여 얻은 결과물이다.

7-1 평균비교

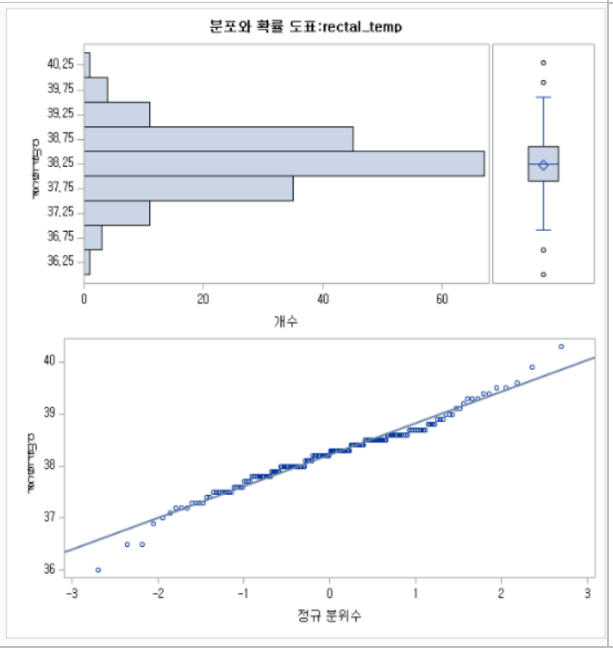
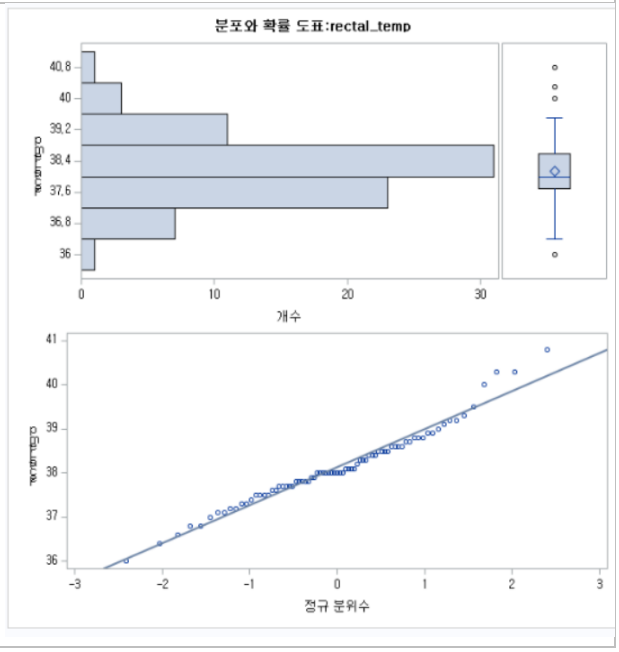
먼저 Rectal_temp에 대한 비교를 가고자 한다. 귀무가설과 대립가설은 다음과 같다.

H0: 두 그룹에 Rectal_temp는 차이가 없다.

H1: not H0

먼저 평균 비교에 대한 가정이 만족하는지에 대한 여부를 살펴 봐야한다. 먼저 각 그룹의 데이터가 서로 독립이라 가정하자. 또한 각 그룹의 데이터는 각각의 평균과 분산을 따르는 정규분포를 따르는 모집단에서 나온 표본이라는 가정이 필요하다. 이를 확인하기 위해 다음과 같은 histogram와 qqplot을 그릴 수 있었다.

<표 4: 생존/사망 그룹에 따른 rectal_temp>

생존한 그룹	사망한 그룹
	

하지만 눈으로 보기에 구분이 어려우므로 좀 더 정밀한 가정검토를 위해 샤피로 윌크 검정을 이용했다. 유의수준은 5%로 정했다.

H0 : 해당 표본이 정규분포를 따르는 모집단에서 나왔다.

H1: not H0

<표 5 : 생존/사망 그룹에 따른 정규성검정>

생존한 그룹

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.977774	Pr < W	0.0060
Kolmogorov-Smirnov	D	0.083294	Pr > D	<0.0100
Cramer-von Mises	W-Sq	0.209359	Pr > W-Sq	<0.0050
Anderson-Darling	A-Sq	1.226517	Pr > A-Sq	<0.0050

사망한 그룹

정규성 검정				
검정	통계량		p 값	
Shapiro-Wilk	W	0.971272	Pr < W	0.0784
Kolmogorov-Smirnov	D	0.098828	Pr > D	0.0626
Cramer-von Mises	W-Sq	0.114266	Pr > W-Sq	0.0750
Anderson-Darling	A-Sq	0.706207	Pr > A-Sq	0.0660

두 그룹 모두 해당 표본이 정규분포를 따르는 모집단에서 나왔다는 가정을 만족하지 못했다. 따라서 해당 데이터에 평균비교를 하는 것은 적절하지 않다. 이는 다른 변수에 대해서도 마찬가지였으며 해당 데이터에 평균비교는 적절치 않다고 판단했다.

7-2 분할표 분석

Outcome을 제외한 이항 반응 값을 가진 surgery, age, capillary_refill_time, surgical_lesion 을 통해 각 변수와 생존 혹은 사망 간의 연관성을 조사해 볼 수 있다. 단 본 데이터는 case-control study 에서 얻어진 데이터이기 때문에 Odds Ratio를 통해 분석 해야 한다.

① Surgery

오즈비 및 상대 리스크			
통계량	값	95% 신뢰한계	
오즈비	2.6670	1.4697	4.8399
상대 리스크(칼럼 1)	1.8897	1.2408	2.8781
상대 리스크(칼럼 2)	0.7085	0.5873	0.8548

빈도
백분율
행 백분율
칼럼 백분율

테이블:outcome * surgery

outcome	surgery 1	surgery 2	합계
1	83 32.55	95 37.25	178 69.80
2	19 7.45 24.68 18.63	58 22.75 75.32 37.91	77 30.20
합계	102 40.00	153 60.00	255 100.00

<그림 3 >

오즈비가 2.6670 이므로 수술을 받지 않았을 때 받았을 때에 비해 생존하는 오즈가 2.6670 배 더 많았다. 또한 95% 신뢰구간이 1을 포함하고 있지 않으므로 이는 5% 유의수준 하에서 유의한 결과였다고 생각할 수 있다.

② age

빈도
백분율
행 백분율
칼럼 백분율

테이블:outcome * age

outcome	age 1	age 2	합계
1	12 4.71 6.74 52.17	166 65.10 93.26 71.55	178 69.80
2	11 4.31 14.29 47.83	66 25.88 85.71 28.45	77 30.20
합계	23 9.02	232 90.98	255 100.00

오즈비 및 상대 리스크

통계량	값	95% 신뢰한계	
오즈비	0.4337	0.1824	1.0316
상대 리스크(칼럼 1)	0.4719	0.2178	1.0225
상대 리스크(칼럼 2)	1.0880	0.9851	1.2017

<그림 4>

오즈비가 0.4337 이므로 생후 6개월 미만이었던 말은 성년인 말에 비해 생존하는 오즈가 0.4337 배 더 적었다. 하지만 95% 신뢰구간이 1을 포함하고 있지 않으므로 이는 5% 유의수준 하에서 유의하지 않은 결과라고 할 수 있다.

③ capillary_reffill_time

빈도
백분율
행 백분율
칼럼 백분율

테이블:outcome * capillary_refill_time

outcome	capillary_refill_time 1	2	합계
1	147 57.65 82.58 77.78	31 12.16 17.42 46.97	178 69.80
2	42 16.47 54.55 22.22	35 13.73 45.45 53.03	77 30.20
합계	189 74.12	66 25.88	255 100.00

오즈비 및 상대 리스크

통계량	값	95% 신뢰한계	
오즈비	3.9516	2.1846	7.1478
상대 리스크(칼럼 1)	1.5140	1.2214	1.8768
상대 리스크(칼럼 2)	0.3831	0.2561	0.5732

<그림 5>

오즈비가 3.9516이므로 capillary refill time 이 3초 미만이었던 경우 3초 이상이었던 경우보다 생존하는 오즈가 3.9516배 더 많았다. 또한 95% 신뢰구간이 1을 포함하고 있지 않으므로,이는 5% 유의수준 하에서 유의한 결과였다고 생각할 수 있다. Capillary refill time 이 길수로 혈류가 원활하지 않다는 것을 의미하므로 예상과 같은 결과라 할 수 있다.

④ surgical_lesion

빈도
백분율
행 백분율
칼럼 백분율

테이블:outcome * surgical_lesion

outcome	surgical_lesion 1	2	합계
1	92 36.08 51.69 57.14	86 33.73 48.31 91.49	178 69.80
2	69 27.06 89.61 42.86	8 3.14 10.39 8.51	77 30.20
합계	161 63.14	94 36.86	255 100.00

오즈비 및 상대 리스크

통계량	값	95% 신뢰한계	
오즈비	0.1240	0.0564	0.2730
상대 리스크(칼럼 1)	0.5768	0.4910	0.6776
상대 리스크(칼럼 2)	4.6503	2.3717	9.1181

<그림 6>

오즈비가 0.1240 이므로 문제가 외과적 병변이었을 때 외과적 병변이 아니었을 때에 비해 생존하

는 오즈가 0.1240 배 더 적었다. 또한 95% 신뢰구간이 1을 포함하고 있지 않으므로 이는 5% 유의수준 하에서 유의한 결과였다고 생각할 수 있다.

⑤ 코크란-맨텔-핸젤 검정

위에서의 분할표 분석에서 age 변수는 오즈비의 95% 신뢰구간이 1을 포함해 유의하지 않게 결론이 나왔다. 하지만 층을 제어할 시에 결과가 달라질 수도 있다고 생각하여 앞에서의 분석을 통해 중요한 변수임을 알 수 있었던 surgical_lession을 층변수로 제어하여 CMH 분석을 하기로 했다.

테이블 outcome * age에 대한 요약 통계량
제어 변수 : surgical_lesion

Cochran-Mantel-Haenszel 통계량 (테이블 스코어에 기반한)

통계량	대립가설	자유도	값	Prob
1	영(0)이 아닌 상관계수	1	3.0983	0.0784
2	행 평균 스코어 차이	1	3.0983	0.0784
3	일반 연관성	1	3.0983	0.0784

공통 오즈비 및 상대 리스크

통계량	방법	값	95% 신뢰한계	
오즈비	Mantel-Haenszel	0.4522	0.1796	1.1381
	로짓	0.4215	0.1703	1.0431
상대 리스크(칼럼 1)	Mantel-Haenszel	0.4913	0.2210	1.0919
	로짓	0.4473	0.2029	0.9860
상대 리스크(칼럼 2)	Mantel-Haenszel	1.0892	0.9756	1.2159
	로짓	1.0746	0.9676	1.1933

<그림 7>

H0 : surgical_lession을 제어한 상태에서 생존여부와 나이간에 상관관계가 없다.

H1 : not H0

코크란-맨텔-헨젤 통계량의 값이 3.0983이며 p-value 는 0.0784 였다. 따라서 유의수준 0.05 하에서 귀무가설을 기각할 충분한 증거가 없다. 따라서 surgical_lesion을 제어한 상태에서 age 변수와 outcome 변수간에는 연관이 없다고 결론 내린다. 이는 맨탤 헨젤의 오즈비의 95% 신뢰구간이 1을 포함하고 있는 것에서도 확인 가능하다.

7-3 로짓 모델

로짓모델을 적합하려 한다. 모든 변수들을 전부 고려할 시 모델의 효용성이 적어질 수 있으므로 stepwise selection 방법을 통해 변수 선택을 실시하였다. 그 과정은 다음과 같았다.

Summary of Stepwise Selection

Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	pain		4	1	71.7094		<.0001
2	packed_cell_volume		1	2	24.6759		<.0001
3	surgical_lesion		1	3	14.1621		0.0002
4	age		1	4	9.9722		0.0016
5	peristalsis		3	5	10.5800		0.0142
6		peristalsis	3	4		6.5359	0.0883

<그림 8>

첫번째 스텝에서 다섯 번째 스텝까지 pain, packed_cell_volume, surgical_Sesion, age, peristalsis 가 차례로 모형에 들어왔다는 것을 확인 할 수 있었다. 하지만 peristalsis의 경우 마지막 여섯 번째 스텝에서 p-value 가 유의수준 0.05 하에서 유의하지 않아 모델에서 빠진 것을 확인 할 수 있었다. 이렇게 적합된 모델이 데이터를 잘 설명해주는지 확인하고자 Goodness of Fit 테스트를 했다.

Deviance and Pearson Goodness-of-Fit Statistics

Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	190.1721	243	0.7826	0.9948
Pearson	215.5830	243	0.8872	0.8967

<그림 9>

H0 : 로짓 모형이 잘 적합된다.

H1 : 로짓 모형이 잘 적합되지 않는다.

적합도 검정에서의 Deviance 와 Pearson 의 검정 통계량이 각각 190.17, 215.58 이었다. P-value

는 0.9948, 0.8967 이었으므로, 유의수준 5% 하에서 귀무가설을 기각할 수 없다. 즉 로짓 모형이 잘 적합 된다고 할 수 있다. 다음은 모델에서의 통계량이다.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	6.5665	1.3022	25.4274	<.0001
age	1	1	-1.7574	0.5816	9.1295	0.0025
pain	1	1	0.8948	0.7594	1.3884	0.2387
pain	2	1	1.7458	0.5255	11.0384	0.0009
pain	3	1	2.1276	0.5557	14.6583	0.0001
pain	4	1	-0.6071	0.5622	1.1661	0.2802
packed_cell_volume		1	-0.1014	0.0203	24.9931	<.0001
surgical_lesion	1	1	-2.2107	0.5934	13.8780	0.0002

<그림 10>

따라서 적합된 모델은 다음과 같다.

$$\ln\left(\frac{P(outcome=1)}{1-P(outcome=1)}\right) = 6.5665 - 1.7574*age + pain\_1 * 0.8948 + 1.7458 * pain\_2 + 2.1276*pain\_3 -0.6071*pain\_4 + -0.1014*packed\_cell\_volume -2.2107*surgical\_lesion$$

대부분의 추정된 계수는 유의수준 5% 하에서 유의하였지만, pain이 1인 수준과 pain이 4 수준인 경우의 계수는 유의수준 5% 하에서 유의하지 않았다. 각 인자들과 outcome 과의 관계는 다음 표에서도 확인 할 수 있다.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
age 1 vs 2	0.172	0.055	0.539
pain 1 vs 5	2.447	0.552	10.838
pain 2 vs 5	5.731	2.046	16.051
pain 3 vs 5	8.395	2.825	24.949
pain 4 vs 5	0.545	0.181	1.640
packed_cell_volume	0.904	0.868	0.940
surgical_lesion 1 vs 2	0.110	0.034	0.351

<그림 11>

생후 6개월 전 말의 경우 성년의 말 보다 생존을 할 오즈는 0.172배 적었다. 또한 pain이 2(depressed) 이였던 수준은 5인 수준(지속적인 심한 고통)에 비해 생존할 오즈가 5.731배 높았으며, pain이 3이었던 경우(간헐적인 약한 고통) 5였던 수준(지속적인 심한 고통)에 비해 생존할 오즈가 8.396배 더 높았다. 또한 외과적 병변 문제였던 경우 아닌 경우보다 생존할 오즈가 0.11 배 더 낮았다. 마지막으로 packed_cell_volume이 한 단위 증가할 때 생존할 오즈가 0.904 배 더 낮아졌다. 이 변수들의 95% 신뢰구간은 모두 1을 포함하고 있지 않다는 것을 확인할 수 있었다. 그러나 pain_1(no pain)과 pain_4(간헐적인 심한 고통)의 경우 95% 신뢰구간이 1을 포함하고 있어 유의하진 않았다.

7-4 비례오즈 로짓모형

추가적으로 temp_of_extremities는 순서형 반응 범주 이므로, 비례오즈 로짓 모형을 통해 분석해 볼 수 있다. 사지의 체온이 차가운 쪽에 가깝다면 쇼크로의 가능성이 있다.

이 때의 가정은 4-1=3 개의 범주에 대해 설명변수들의 효과가 동일하다는 것이다. 이 가정이 충족하는가를 알아보기 위한 테스트 결과는 다음과 같다.

Score Test for the Proportional Odds Assumption		
Chi-Square	DF	Pr > ChiSq
3.9880	6	0.6783

<그림 12>

H0 : 비례오즈가정이 만족한다.

H1 : 비례오즈가정이 만족하지 않는다.

카이제곱 통계량 값은 3.9880 이고 p-value 는 0.6783 이므로 유의수준 0.05 하에서 귀무가설을 기각할 충분한 증거가 없다. 비례오즈가정이 만족하고 있음을 알 수 있다.

Deviance and Pearson Goodness-of-Fit Statistics				
Criterion	Value	DF	Value/DF	Pr > ChiSq
Deviance	233.9258	261	0.8963	0.8848
Pearson	275.2799	261	1.0547	0.2601

<그림 13>

다음으로 모형의 적합도를 보면 우도비 카이제곱 통계량과 피어슨의 카이제곱 검정 통계량 모두 p-value가 0.05보다 크므로 모형이 자료에 잘 적합된다고 할 수 있다.

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	53.5971	3	<.0001
Score	49.6487	3	<.0001
Wald	47.8885	3	<.0001

<그림 14>

H0 : 모든 인자가 생존/사망 여부에 연관이 없다.

H1 : not H0

우도비 카이스퀘어 통계량이 53.5971 이고 p-value<0.0001 이므로 유의수준 5%하에서 귀무가설을 기각하지 않는다. 따라서 적어도 하나의 인자가 생존/사망 여부에 연관성이 있다는 사실을 알 수 있다.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	1	-6.2572	0.6994	80.0329	<.0001
Intercept	2	1	-3.6467	0.6075	36.0395	<.0001
Intercept	3	1	-1.6142	0.5796	7.7555	0.0054
surgical_lesion	1	1	1.0917	0.2560	18.1816	<.0001
age	1	1	0.6350	0.4210	2.2750	0.1315
packed_cell_volume		1	0.0638	0.0127	25.0793	<.0001

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
surgical_lesion 1 vs 2	2.979	1.804	4.921
age 1 vs 2	1.887	0.827	4.307
packed_cell_volume	1.066	1.040	1.093

<그림 15>

위의 모수 추정치를 바탕으로 비례오즈 로짓모형 식은 다음과 같이 표현될 수 있다.

Logit[F_{i1}] = -6.2572 + 1.0917*surgical_lesion + 0.6350*age + 0.0638*packed_cell_volume, j=1

Logit[F_{i2}] = -3.6467 + 1.0917*surgical_lesion + 0.6350*age + 0.0638*packed_cell_volume,, j=2

Logit[F_{i3}] = -1.6142 + 1.0917*surgical_lesion + 0.6350*age + 0.0638*packed_cell_volume,, j=3

(j 가 높을수록 따뜻한 범주이다)

surgical_lesion 의 계수는 1.0917, age의 계수는 0.6350이며, packed_cell_volume의 계수는 0.0638이다. surgical_lesion의 p-value는 0.0001 이하인데 유의수준 0.05 하에서 사지의 온도 수준에 유의한영향을 주고 있다고 말할 수 있다. 또한 packed_cell_volume의 p-value 역시 0.0001 이하이므로 사지의 온도 수준에 유의한 영향을 주고 있다. 반면 age의 p-value는 0.1315으로 유의수준 0.05하에서 사지의 온도 수준에 유의미한 영향을 주고 있다고 말할 수 없다.

다른 변수가 고정 되어있을 때 surgical_lesion 따른 오즈비는 약 exp(1.0817) =2.979 이다. 따라서 외과적 병변일 경우 아닐 경우보다 사지온도가 차가운 범주에 속할 오즈가 2.979배 증가한다.

다른 변수가 고정되었을 때 age에 따른 오즈비는 약 exp(0.6350) = 1.887 이다. 따라서 생후 6개월 이하의 말이 성년의 말 보다 사지의 온도가 차가운 쪽 범주에 속할 오즈가 1.887 배 증가한다. 하지만 유의한 수준은 아니었다.

다른 변수가 고정되었을 때 packed_cell_volume에 따른 오즈비는 약 exp(0.0638) = 1.066 이다. 따라서 packed_cell_volume이 1단위 증가할 때 더 차가운 쪽 범주에 속할 오즈가 1.066배 증가한다.

7-5 로그-리니어 모델

다음으로 로그-선형 모형을 통해 각 변수들간의 유의한 관계를 알아보도록 하겠다. 변수들은 surgical_lesion과 outcome 그리고 mucous_membrane만을 고려하였다. 가장 적합한 모델을 고르기 위해 다음과 같이 계층적으로 모형을 적합시켰다.

<표 6 : 포화 모델과의 비교>

모형	G^2(df)	p-value
포화모형		
surgical_lesion|outcome outcome|mucous_membrane surgical_lesion|mucous_membrane	0.55(3)	0.9077

H0 : 축소모형이 적합하다. H1: 포화 모형이 적합하다.

3요인 교호작용을 제외한 부분 연관 모형의 우도비 통계량이 0.55 이므로 이는 크리티컬 밸류인 7.81 보다 작으므로 유의수준 0.05 하에서 귀무가설을 기각하지 않는다. 따라서 3요인 교호작용을 제외시킨 모형이 포화모형보다 더 적절하다. 다음으로 나머지 축소 모형을 고려해 보도록 한다.

<표 7 : 부분연관 모형과의 비교>

모형	G^2(df)	G^2((M2|M1)(df)	X^2(0.05,df)
surgical_lesion|outcome outcome|mucous_membrane surgical_lesion|mucous_membrane	0.55(3)		
surgical_lesion|outcome outcome|mucous_membrane	30.28(8)	29.73(5)	11.07
surgical_lesion|outcome surgical_lesion|mucous_membrane	18.16(5)	17.61(2)	5.99
outcome|mucous_membrane surgical_lesion|mucous_membrane	5.92(4)	5.37(1)	3.84

H0 : 축소모형이 적합하다. H1: 확대 모형이 적합하다.

모든 축소모형의 우도비 통계량은 해당 크리티컬 밸류 (유의수준은 5%) 보다 그 값이 컸다. 따라서 귀무가설을 기각할 충분한 증거가 있으므로 귀무가설을 기각하고 확대 모형이 적합하다고 결

론을 내린다. 다시 말해 3요인 교호작용을 제거한 부분 연관모형이 가장 적합한 로그-리니어 모델이라고 말할 수 있다. 적합 시킨 결과는 다음과 같다.

Analysis of Maximum Likelihood Estimates

Parameter		Estimate	Standard Error	Chi-Square	Pr > ChiSq
surgical_lesion	2	-0.5131	0.1131	20.58	<0001
outcome	2	-0.3368	0.1208	7.77	0.0053
surgical_lesion*outcome	2 2	-0.2848	0.1267	5.06	0.0245
mucous_membrane	2	0.2205	0.1754	1.58	0.2089
	3	0.6448	0.1805	12.76	0.0004
	6	-0.5497	0.2388	5.30	0.0214
	1	0.6430	0.2103	9.35	0.0022
	4	-0.3157	0.2666	1.40	0.2365
outcome*mucous_membrane	2 2	-0.0181	0.1764	0.01	0.9182
	2 3	-0.1793	0.1574	1.30	0.2547
	2 6	0.3236	0.2191	2.18	0.1397
	2 1	-0.6812	0.2074	10.79	0.0010
	2 4	0.0157	0.1567	0.01	0.9200
surgical_le*mucous_membr	2 2	0.3371	0.1688	3.99	0.0458
	2 3	-0.00303	0.1700	0.00	0.9858
	2 6	0.0317	0.2452	0.02	0.8972
	2 1	0.5348	0.1510	12.54	0.0004
	2 4	-0.8427	0.2643	10.17	0.0014

<그림 16>

추정된 우도비 추정치를 통해 유의수준 5% 하에서 여러가지 해석을 할 수 있다. 먼저 주어진 데이터에서는 외과적 병변인 경우가 유의미하게 많았다. 또한 생존한 경우가 사망한 경우보다 유의하게 많았다. 한편 주어진 데이터에서 점막유형 3(pale pink)과 1(normal pink) 이 유의하게 많았으며 유형 6(dark cyanotic)은 유의하게 적었다. 하지만 로그-리니어 분석에서 중요한 것은 교호작용이다. 먼저 외과적 병변이 아니면서 사망한 말이 유의하게 적었다고 할 수 있다. 또한 생존하였으며 점막이 유형 1(normal pink) 이었던 경우가 유의하게 많았다. 마지막으로 외과적 병변이 아니면서 점막이 유형 1(normal pink), 유형 2(bright pink) 경우가 많았으며, 외과적 병변이 아니면서 점막이 유형 4(pale cyanotic)인 경우가 유의하게 적었다는 것을 알 수 있었다.

8. 결론

본 프로젝트에서 수행한 여러 통계적 모델과 검증을 통해 말의 생사여부에 미치는 다양한 변수들의 영향에 대해 알아볼 수 있었다. 지금까지 알아낸 내용을 바탕으로 대략적인 정리를 다음과 같이 할 수 있다.

<표 : 8 변수에 대해 알아낸 사실에 대한 요약>

변수	설명
surgery	수술을 받지 않았을 때 생존을 하는 경향이 있고, 수술을 받았을 때 사망하는 경향이 있다.
age	생후 6개월 미만의 말인 경우 성년의 말 보다 더 쉽게 죽는 경향이 있었다. 그러나 그 차이가 유의하지 않은 경우가 대다수이기 때문에 좀 더 조사해야 한다.
Rectal_temp	유의한 관계를 밝혀내진 못했다.
pulse	맥박이 높을수록 사망하는 축에 속할 가능성이 많았다고 해석할 수 있다. 트리 알고리듬과 랜덤포레스트에서 중요한 변수였다.
Respiratory_rate	말의 생존여부에 대해 중요한 변수인지 알 수 없다. (공변량을 고려한 잠재층 모델과, 로지스틱회귀에서의 p-value가 유의수준 0.05보다 높았다.) 변동이 심하기 때문에 이에 대해 통계적 검증을 하는 것도 쉽지 않을 것이라 예상한다.
Temp_of_extremities	사지의 온도가 따뜻한 범주에 속할 때 생존하는 경우가 많았으며, 차가운 범주에 속할 때 사망하는 경향이 있었다. 트리 알고리듬에서 중요한 역할을 했다.
Peripheral_pulse	맥박이 normal 범주인 경우 살아남는 경우가 많았고, reduced 이거나 absent인 경우 사망하는 경우가 많았다.
Mucous_membrane	점막이 pink 계열일 경우 살아남는 경향이 있었고 cyanotic의 경우 죽는 경우가 많았다. 랜덤포레스트에서 중요한 역할을 하는 편이었다.
Capillary_refill_time	모세혈관 충전시간이 3초 이하일 때 살아남는 경향이 있었고 3초 이상일때 죽는 경향이 있었다.
pain	고통의 정도가 심각한 경우 죽는 경향이 있었다. 트리 알고리듬과 랜덤포레스트에서 가장 중요한 변수였다.
peristalsis	다중대응분석에 의하면 연동운동이 과운동성인 경우가 좀더 살아남는 경향이 있었고 저 운동성인 경우 죽는 경향이 있었다. 그러나 잠재층 분

	석에서 생존할 것으로 추정되는 Class1에서 저운동성과 과운동성 모두 많아 판단을 내릴 수 없었다.
Packed_cell_volume	혈액 속 적혈구 수가 늘어날수록 사망하는 경향이 있었다. 랜덤포레스트에서 중요한 역할을 하는 변수였다.
Surgical_lesion	외과적 병변이 아닌 경우 생존하는 경향이 있었으며, 외과적 병변인 경우 사망하는 경향이 있었다.

여기서 주목해야 할 점은 각 설명변수들이 독립적으로 말의 생존여부에 영향을 미치는 것이 아니라 서로 연관이 있다는 점이다. 예컨데, 말이 생존하지 못하는 경우에 대해 생각해보면, 외과적 병변일 경우(당연히 surgery 변수와 연관이 있을 것이다.) 혈류가 좋지 않을 것이며, 이때 모세혈관 충전시간은 늘어날 것이며, 혈액 속 적혈구 수 또한 늘어날 것이다. 또한 주변 맥박 또한 reduced 나 absent 에 속하게 될 것이며, 점막은 pink 계열에서 멀어지고 cyanotic 계열일 가능성이 높아질 것이다. 또한 혈류가 좋지 못하면 사지의 온도가 내려가는 것은 상식이다. 따라서 이 변수들은 모두 혈류 때문에 variation 이 생긴다고 볼 수 있다. 트리 알고리듬에서 사용한 최선의 변수의 수가 단 3개 (pain, pulse, temp_of_extremiteis) 로 총 변수 수에 비해 적었던 것 또한 이러한 이유일 것에서 일 것이라 생각한다.

본 분석에서의 한계는 결측치 대체에 대한 미흡함이다. 본 프로젝트에서 사용한 결측치 대체 방법은 MAR 가정을 만족해야 적절하다고 할 수 있지만, 도메인 지식 과 통계적 지식 부족으로 본 데이터에서 MAR를 만족했는지 알 수 없었다. 또한 본래는 여러 개의 imputation을 한 complete 데이터 셋 각각에 대해 분석을 한 다음에 이들 결과를 종합해야 하지만 여기서는 한 개한 complete 데이터 셋 만을 사용했으므로 적절한 결측 데이터 분석을 했다고 말할 수는 없다.

또한 결측치가 많은 여러 설명변수들을 데이터 분석에 고려하지 않은 것 또한 문제를 일으켰다고 생각한다. 앞에서 언급한 바와 같이 많은 변수의 경우 혈류에 직접적인 영향을 받는 변수들이다. 즉 데이터가 반복되어 설명되고 있는 상황이었으며, 혈류와는 관계가 상대적으로 적은 여러 변수들을 충분히 고려하지 못했다고 생각한다. (모델링을 통한 생존여부 예측의 정분류율이 70%대로 낮은 것의 원인이었을 것이라 생각한다.) 따라서 이들 변수를 포함시킨다면 정 분류율이 올라갈 수 있을 것이라 예상한다.

하지만 그렇다 하더라도 본 프로젝트의 다양한 분석을 통해, 대략적인 변수들간의 관계부터 엄밀한 통계적 분석방법을 통한 유의성 검정까지 다방면에서 말의 생사에 영향을 줄 수 있는 변수들을 설명하고 해석할 수 있어 의미 있는 분석이었다고 생각한다.

참고문헌

이재원/박미라/유한나(2005).「생명과학연구를 위한 통계적 방법」. 자유아카데미.

허명회(2014).「응용데이터분석 Applied Data Analysis Using R」. 자유아카데미.

데이터 출처: Horse Colic Dataset, Kaggle. (https://www.kaggle.com/uciml/horse-colic)